RegEMT 👑 Regressive Ensemble for Machine Translation Quality Evaluation

Michal Štefánik and Vít Novotný and Petr Sojka

Faculty of Informatics, Masaryk University, Brno, Czechia

1. Problem

Src: "Ich habe diesen Artikel nie geschrieben, ich habe ihn nur bearbeitet."

Ref: "I never wrote this article, I just edited it."

	[1] BERTScr(Ref1, H)	[2] BLEUrt(Ref1, H)	[3] Prism(Ref1, H)	[4] Comet(Ref1, H)	Compositionality*(Ref1, H)
H1: "I wrote the article, but I never edited it."	0.939	0.499	- 2.145	0.721	0.267
H2: "Although I did not write the article, I edited it."	0.933	0.479	- 1.601	0.675	0.282

2. Our approach

Idea: Eliminate single metric-specific failures by robustly relying on multiple language properties.

RegEMT aims to **ensemble** a variety of measures reflecting *wider* range of **levels of language similarity,** covering **blind spots** of deep semantic metrics**:**

1. Surface-level	BLEU, METEOR
2. Syntactic level	Compositionality*
3. Word-level	Embeddings: {FastText* / Bert}
	X matching: {Max π / Soft Cos* / WMD*}
4. Semantic	PRISM, Comet, BLEUrt

* measures that were either developed, or used for evaluating machine translation quality for the first time in our work.

RQ1: Can an ensemble of surface, syntactic, and semantic-level metrics significantly improve the performance of single metrics?

RQ2: Can such an approach be applied cross-lingually, i.e., on languages that it has not been trained on?

RQ3: Can surface-level metrics in reference-free configuration achieve results comparable to the reference-based ones?

RQ4: Are contextual token representations important for evaluating semantic equivalence, or can these be replaced with *cheaper* (pre-inferred and averaged) token representations?

3. Results

	RegEMT 👑	Prism	BERTScr	WMD-cont	WMD-dec	WMD-dec-tf	SCM-dec	SCM-dec-tf	Compos	Reg-base	Comet	SCM	SCM-tf	WMD	WMD-tf	BLEUrt	BLEU	METEOR
MQM-src ***zh-en***	**.59**	.36	.44	.44	.29	.17	.19	.13	.13	.34								
MQM-src ***zh-en-X***	**.49**	.36	.44	.44	.29	.17	.19	.13	.13	.34								
MQM-src ***en-de***	**.36**	.09	.14	.06	.04	.07	.03	.02	.23	.28								
MQM-src ***en-de-X***	**.31**	.09	.14	.06	.04	.07	.04	.02	.23	.28								
MQM-ref ***zh-en***	**.62**	.45	.45	.43	.27	.21	.10	.26	.01	.35	.51	.19	.35	.29	.27	.48	.25	.28
MQM-ref ***zh-en-X***	**.62**	.45	.45	.43	.27	.21	.09	.25	.01	.31	.51	.19	.35	.29	.27	.48	.25	.28
MQM-ref ***en-de***	**.60**	.32	.22	.25	.32	.28	.33	.18	.12	.27	.48	.06	.14	.13	.07	.10	.13	.20
MQM-ref ***en-de-X***	.38	.32	.22	.25	.32	.28	.34	.17	.12	.29	**.48**	.06	.14	.13	.07	.10	.13	.20
DA 2016-src	**.84**	.72	.74	.73	.57	.51	.37	.51	.29	.18	.82	.39	.45	.44	.42	.81	.42	.50
DA 2016-tgt	.68	**.70**	.34	.25	.09	.10	.10	.24	.13	.04								
catastrophic-src	**.29**	.26	.13	.11	.12	.15	.13	.09	.10	.09								

Table 1: **(RQ1)** Spearman's correlation of the ensemble (*RegEMT*) and standalone ensembled metrics with expert judgements (MQM), direct assessments (DA, averaged) and catastrophic errors (MLQE-PE data set, averaged).
Cross-lingual results are suffixed with **X (RQ2).**
(RQ3) A surface-level regressor using only source Wordpiece length of source and target (*Reg-base*) consistently outperforms BLEU and METEOR and reaches competitive results even in reference-free evaluation (*src*).

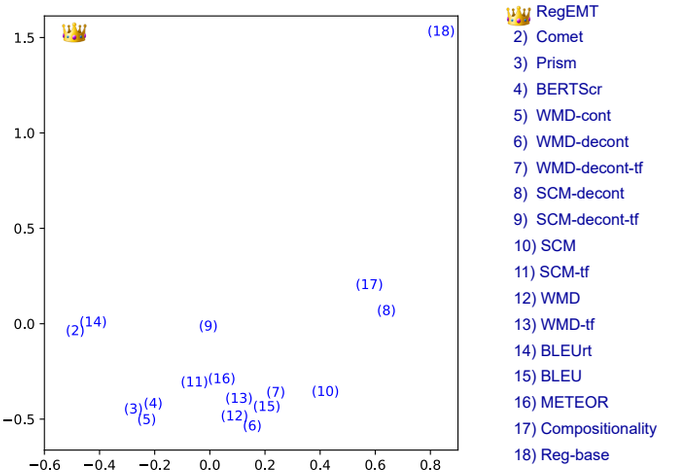

Figure: Two-dimensional PCA projection of the ensembled metrics by mutual correlation of their values, evaluated on *zh–en* language pair of MQM. We can see a high correlation of *contextualized embedding metrics* (3, 4, 5), as well as *non-contextual* ones (6 –13), which also correlate high with *surface-level* measures (15, 16). Ensemble measures are the most orthogonal to the others.

	WMD-cont	WMD-dec	Δ
MQM-src ***zh-en***	.44	.29	− .15
MQM-src ***en-de***	.06	.04	− .02
MQM-ref ***zh-en***	.43	.27	− .16
MQM-ref ***en-de***	.25	.32	+ .07

← **Table 2**: **(RQ4)** Impact of *decontextualization* by averaging pre-inferred token representations to a correlation of WMD with MQM is significant, but not consistent.
A computation of *SCM-cont* is no longer memory-feasible.

4. Takeaways

1. Using an ensemble of measures covering specific defects is more *interpretable* and can be an easier future direction than searching for a single, omnipotent training objective.
2. Simple regressor trained on *Wordpiece* text lengths can beat commonly-used surface metrics of BLEU and METEOR.
3. Results show that ensemble approach can *extrapolate* to other languages, keeping the usability of multilingual measures.

References

👑 Štefánik, M., Novotný, V., Sojka, P. (2021). Regressive Ensemble for Machine Translation Quality Evaluation. WMT.
[1] Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. ICLR.
[2] Sellam, T., Das, D., & Parikh, A.P. (2020). BLEURT: Learning Robust Metrics for Text Generation. ACL.
[3] Thompson, B., & Post, M. (2020). Automatic Machine Translation Evaluation in Many Languages via Zero-Shot Paraphrasing. EMNLP.
[4] Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Çelikyilmaz, A., & Choi, Y. (2019). COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. ACL.

Corresponding author: Michal Štefánik, stefanik.m@mail.muni.cz